EDA - Cheat Sheet

Exploratory Data Analysis

EDA is a process of analyzing the datasets to summarize their main features using numerical and visual methods.

Data:

Data is units of information in a structured or unstructured format.

Numerical Variable:

A continuous numeric variable is a variable that can have an infinite number of values within a range.

Categorical Variable:

The categorical variable has two or more levels
Also known as a qualitative variable.

Nominal Data:

Nominal data has no order and has two or more two categories.

Binary Data:

Binary data has no order and has strictly two categories.

Ordinal Data:

Ordinal Data is ordered nominal data.

Descriptive statistics:

- Descriptive statistics is the term given to the analysis of the data that helps describe, visualize or summarize the data.
- Helps exhibit the patterns in the data

Mean:

The mean is the sum of all observations divided by the number of observations. It is an average of data.

$$\bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$$

Median:

Median is the middle most observation in the data when it is arranged either in ascending or descending order of their values.
It divides the data into two equal parts. Thus, it is a positional average.

dataframe['column'].median()

Mode:

The mode of the data is the value that has the highest frequency. In simple words, it is the most repeated observation.

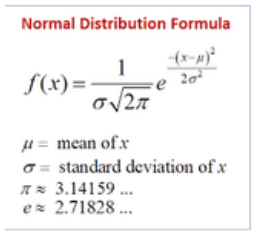
dataframe['column'].mode()[0]

Distribution of the Data:

The distribution is a summary of the frequency of values taken by a variable.
The distribution of the data gives information on the shape and spread of the data.

Normal distribution:

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



Measures of Dispersion:

The measure of dispersion refers to the variability within the data.
Variability is the measure of how close or far the data lie from the central value.

Variance:

Variance is the arithmetic mean of squares of deviations taken from the mean.

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

Standard deviation:

The standard deviation of the variable is the square root of the variance.

$$\sigma = \sqrt{\frac{\Sigma\,(x-\mu)^2}{N}}$$

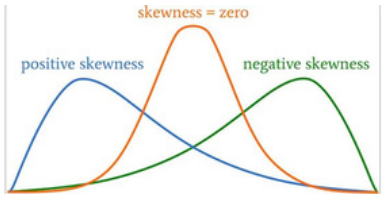
Coefficient of variation:

The coefficient of variation is a statistical measure of the dispersion of data points around the mean.

$$CoV = \frac{Standard\ Deviation}{Mean}$$

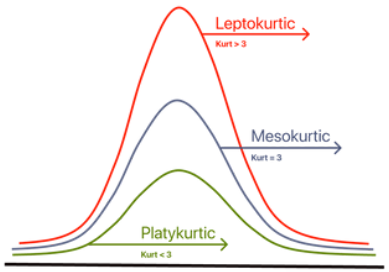
Skewness:

Skewness is a lack of symmetry or departure from symmetry.
If the distribution of the data is elongated on either side then the data is said to be skewed.



Kurtosis:

Kurtosis measures the peakedness of the distribution. it is a statistical measure that defines how the tails of the distribution differ from the normal distribution



Covariance:

Covariance is a measure of how much two random variables vary together. it explains how two variables vary together.

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Correlation:
Correlation explains how one or more variables are related to each other.

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

Encoding:
It is a mathod of converting categorical variables into numerical values.

Few types of encoding:

(N-1) dummy encoding: It is used to create dummy variables from a categorical variable. If a categorical variable has 3 values like male, female, and transgender it will create 2 variables (k-1).

syntax: note: let's assume df is the data frame

pandas.get_dummies(df['column'],drop_first = True)

One hot encoding: It is used to create dummy variables from a categorical variable.
Each category is converted into one column with values '0' and '1', depending on the presence or absence of the category in the corresponding observation.

syntax:

pandas.get_dummies(df['column'])

Label encoding: The LabelEncoder considers the levels in a categorical variable in alphabetical order for encoding. if the variable has 3values like male, female, and transgender. then the converted variable has 0,1,2. 0 for female, 1 for male, and 2 is transgender.

syntax:

from sklearn.preprocessing import LabelEncoder
df['column'] = LabelEncoder().fit_transform(df['column'])

Frequency encoding: Frequency encoding replaces each label of the categorical variable by the percentage of observations within that category.

Syntax:

freq = df.groupby('column').size()/len(df)
df['column'] = df['column'].map(freq)

Z-score

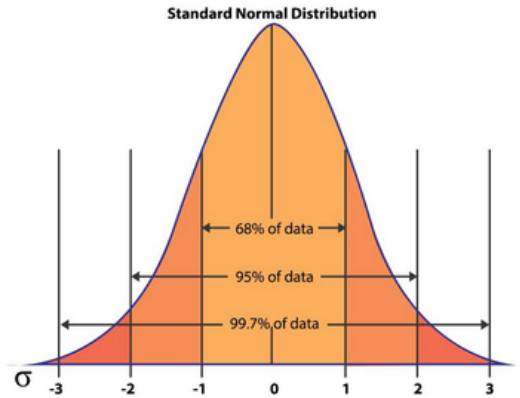
Z-score indicates how much a given value differs from the standard deviation. The z-scores are extensively used in statistics, especially in hypothesis testing.

$$Z = \frac{x - \mu}{\sigma}$$

Standard Normal Distribution:
It is a normal distribution with a mean of zero and a standard deviation of 1.

Empirical Rule: It is also called as 3 sigma rule.
It states when the data is normal distribution then

- 68% of the values lie within the 1 std of the Mean
- 95% of the values lie within the 2 std of the mean.
- 99.7% of the values lie within the 3 std of the mean



Feature scaling: Feature scaling is also known as data normalization. It is a technique used to transform the data into a common scale.

Standardization or Z-score Normalization:
Standardization transforms the data such that the data has a mean 0 and unit variance.

syntax: from sklearn.preprocessing import StandardScaler
df['column'] = StandardScaler().fit_transform(df['column'])

Min-max normalization: Performs linear transformation on the original data. values between 0 to 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Data Transformation

Data transformation is the process of converting data from one format to another.

Log transformation:
Reduces the skewness in the distribution of the original data. It converts exponential growth to linear growth.

Syntax: from numpy import log
log(df['column']

Exponential transformation:
It is used to convert the log-transformed values to their original units. It's just the inverse of log transform.

Syntax: from numpy import exp
exp(df['column'])

Box-cox transformation:

- The generalized version of log transformation.
- Reduces skewness of the data making it more symmetrical

$$\begin{cases} y = \frac{x^\lambda - 1}{\lambda} \text{ where } \lambda \neq 0 \\ y = \ln x \text{ where } \lambda = 0 \end{cases}$$

Syntax:

import scipy.stats as sts
sts.boxcox(df['column'])

Yeo-Johnson Transformation:
Box-Cox Transformation has a limitation it works strictly only on Positive Values so when the data has Negative Values box-cox technique will fail so we can use Yeo-Johnson transformation.

$$\psi(\lambda, y) = \begin{cases} ((y+1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1)]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

Syntax:

from scipy import stats as sts
sts.yeojohnson(df['column']

Data types and Missing values

variable type:

pandas **dtypes** method is used to identify the variable(data) types. It is necessary to make sure all variables have correct data types according to their domain.

- **The syntax for checking the variable types:**

let's say df is a data frame with n variables.

```
df.dtypes
```

- **The syntax for changing the variable data type:**

```
df['column'].astype('int')
```

Missing values:

It is defined as the data value that is not stored for a variable in the observation of interest.

Two types of missing values:

- **Standard Missing values:**

Standard missing values are the values that pandas can detect.

Syntax:

df.isnull() will return True if the missing valuer is present for a particular cell or else return False.

df.isnull().sum() will return total number of missing values for each attribute (columns).

- **Non-standard missing values:**

Sometimes the missing values have different formats. ex:? mark. In this case, Pandas would not identify those values as missing values.

To convert them as missing values manually convert the "?" mark as missing values.

example:

```
df = df.replace("?",np.nan)
```

Handling Missing values:

- **Impute the missing values:** We can impute missing values with one best parameter based on the distribution.
- Drop the missing values: Removing the row with the missing values will lead to loss of information.

Outliers and Treatment

Outliers:

Outliers are the values that look different from the other values in the data.

Discover outliers' presence:

- Using visualization methods like distribution plots.
- Z score method.
- Skewness and Kurtosis.
- And Comparing Mean and Median.

Handling Outliers:

Zscore:

Z-scores are the number of standard deviations above or below the mean.

In most cases: if the z-score values above +3 and below - 3 values will be treated as outliers. This would vary based on industry to industry.

Interquartile range:

The interquartile range is the middle 50% of the dataset.

- It ranges between the third and the first quartile.
- We use the interquartile range, first quartile, and third quartile to identify the outliers.

The outlier is a point that falls below Q1 – 1.5×IQR or above Q3 + 1.5×IQR.

Syntax to remove the outliers:

```
df = df[~((df['column] <(Q1 - 1.5 * IQR)) | (df[column] > (Q3 + 1.5 * IQR)))]
```

Train Test Split:

This will be used when you want to build and evaluate model performance.

Usually, We split the data into two parts and train and test with a common ratio of 70:30 but It depends on the datasets and business.

Syntax:

```
from sklearn.model_selection import train_test_split
X = df[['predictors']], y = df['target_variable']
X_train,X_test,y_train,y_test =
    train_test_split(X,y,test_size = 0.3)
```

Uni-Bi-Multi variate analysis

Univariate Analysis:

- Univariate analysis is the simplest form of statistical analysis.
- It is the study of a single (unit) variable.

Few important types of Univariate analysis:

For Numerical:

1. Summary statistics
2. Histogram
3. Box plot, etc

For categorical:

1. Summary statistic.
2. frequency table and bar plot.

bivariate analysis:

- Understand the relationship between two variables.
- Many businesses, marketing, and social science problems could be solved by using bivariate analysis.

Few methods related to Bivariate analysis are:

1. Line plot, Scatter plot.
2. Box plot, violin plot.
3. Cross tab, Stacked bar plot(count plot using hue).

Multivariate analysis:

- It is the analysis of two or more variables
- Used to study the relationships between the variables
- Useful in determining the effect of one variable on other variables

Few methods related to Multivariate analysis are:

1. Scatter plot using hue.
2. Correlation matric.
3. Heat map of correlation
4. Pair plot